Expert Systems with Applications 40 (2013) 27–33

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

SumView: A Web-based engine for summarizing product reviews and customer opinions

Dingding Wang [a], Shenghuo Zhu [b], Tao Li [a,*]

[a] School of Computing and Information Sciences, Florida International University, Miami, FL 33199, United States
[b] NEC Laboratories America Inc., 10080 N. Wolfe Rd., SW3-350, Cupertino, CA 95014, United States

ARTICLE INFO

Keywords:
Product review
Opinion mining
Summarization
Matrix factorization

ABSTRACT

In this paper, we develop SumView, a Web-based review summarization system, to automatically extract the most representative expressions and customer opinions in the reviews on various product features. Different from existing review analysis which makes more efforts on sentiment classification and opinion mining, our system mainly focuses on summarization, i.e., delivering the majority of information contained in the review documents by selecting the most representative review sentences for each extracted product feature. Comprehensive case studies and experiments demonstrate the effectiveness of our system, and the user study shows users' satisfaction.

1. Introduction

With the rapid growth of e-business, the Web has provided an excellent platform for business to consumer (B2C) electronic commerce (Gamon, Aue, Corston-Oliver, & Ringger, 2005; Jin, Ho, & Srihari, 2009). And Web-based product review systems provide valuable customer feedback to both the merchants who want to keep track of customer opinions on their products and the potential customers who are making informed decisions on whether to purchase the products (Hu & Liu, 2004). On one hand, merchants selling products on the Web often ask their customers to review the products that they have purchased and the associated services. On the other hand, in addition to comparing product specifications for better purchasing decision, consumers now typically read product reviews to identify the best products that fit their preferences. As E-commerce is becoming more and more popular, the repository of customer reviews that a product receives grows rapidly (Archak, Ghose, & Ipeirotis, 2007; Miao, Li, & Dai, 2008). For many popular products, the number of reviews is usually over hundreds or even thousands. This vast richness of content has made it difficult for a user to read through the reviews one by one and to extract useful information such as product qualities and services (Cheung, Kwok, Law, & Tsui, 2003; Pang, Lee, & Vaithyanathan, 2002; Popescu, Nguyen, & Etzioni, 2005; Tellis & Johnson, 2007).

Most of the existing work on mining customer reviews focuses on opinion feature extraction and adjective orientation identification. For example, Hu and Liu Hu and Liu (2004) proposed a review summarization system to extract the opinion noun words using association mining and determine the orientation of the nearby adjective words using the information of synonyms and antonyms in WordNet (Miller, 1995). Their system finally lists all the positive sentences and negative sentences with respect to each product feature. Many recent efforts have been conducted to improve the accuracy of feature extraction (Somprasertsri & Lalitrojwong, 2008) and sentiment analysis (Cui, Mittal, & Datar, 2006). However, there are still some limitations of these existing review summarization systems. First, the accuracy of current automatic feature extraction methods is low (around 0.5–0.7 in precision). It is thus impractical to use in real applications. Second, identification of the opinion words orientation is not satisfactory. As a consequence, the results of statistics on positive and negative opinions are not reliable. Third, current systems usually present their results as a list of opinion sentences or terms. When there are a large number of such sentences or terms, further organization, exploration or summarization processes are still needed.

To address the above issues, we propose SumView, a Web-based review summarization system, to automatically extract the most representative expressions and customer opinions in the reviews on various product features. A prototype of the system can be found at http://rev-sum.appspot.com/. Different from many other systems which use benchmark datasets, SumView is a real Web-based system integrating review crawling from Amazon.com, automatic product feature extraction along with a text field where users can input their desired features, and sentence selection using the proposed feature-based weighted non-negative matrix factorization algorithm. Finally, the most representative sentences are selected to form the summary for each product feature.

* Corresponding author.
E-mail addresses: dwang003@cs.fiu.edu (D. Wang), zsh@nec-labs.com (S. Zhu), taoli@cs.fiu.edu (T. Li).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.070

The rest of this paper is organized as follows. Section 2 discusses the related work on current customer review mining techniques and multi-document summarization. Section 3 introduces the framework of our SumView system. The methods used in SumView are described in Section 4. Illustrative case studies and a user study are presented in Sections 5 and 6 respectively. Finally Section 7 concludes.

2. Related work

2.1. Mining and summarizing customer reviews

In general, mining and summarizing customer reviews involve three tasks: feature identification, sentiment analysis, and summarization. Specifically, feature identification aims to important product features; sentiment analysis is to identify the polarity of the opinions expressed on the features; and summarization aims to deliver the condensed results to users.

2.1.1. Opinion feature extraction

Hu and Liu Hu and Liu (2004) propose an associate mining based method to identify product features. In their work, NLProcessor linguistic parser is first used to parse each review and produce the part-of-speech tag for each word. Then association miner is applied to the transactions of noun/noun phrases to discover frequent features. After two types of pruning, nouns/noun phrases with nearby adjective words from the frequent features are identified as the opinion features. Yi and Niblack Yi and Niblack (2005) define a set of feature term extraction heuristics and select feature terms from the noun phrases obtained from the review texts based on a given topic. Popescu and Etzioni Popescu et al. (2005) propose an unsupervised information extraction system to obtain candidate frequent nouns by setting a frequency threshold. The candidates are then evaluated by computing the mutual information between a candidate and a product class. A more recent work (Meng & Wang, 2009) clusters multiple specifications to extend the vocabulary of product features, and the hierarchical structure is also constructed from the specifications to assist the feature extraction.

2.1.2. Sentiment analysis

Sentiment analysis is usually involved to identify if an opinion feature or sentence is positive or negative. Shallow analysis methods (e.g. Hu & Liu (2004)) typically use a set of seed adjectives which are labeled manually, and then propagate the partial label information on a term graph constructed using external sources such as WordNet. Sentiment analysis can be naturally treated as a two-class classification problem. Machine learning techniques, such as Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM), are widely used to classify the review texts into positive or negative categories (Cui et al., 2006; Pang et al., 2002). Recently matrix factorization techniques have also been used to perform sentiment analysis with lexical prior knowledge (Li, Zhang, & Sindhwani, 2009). Our system does not incorporate sentiment analysis and focuses on summarization.

2.1.3. Result summarization

Current work on result summarization usually list and count all the opinion sentences. For example, Hu and Liu (2004) returns all the positive and negative sentences for each extracted product feature, and a count is provided to illustrate the number of positive and negative opinions for each feature. (Carenini, Ng, & Pauls, 2006; Gamon et al., 2005) use a tree map to visualize product features and the corresponding sentiment assignments associated with the features. These result presentations are efficient in many scenarios, however, users may still need to read through all the opinion sentences or phrases. (Meng & Wang, 2009) reports several most frequent terms or phrases as the summary of a product feature. And Lu, Zhai, and Sundaresan (2009) provides a view of aspect ratings for each product. However, summaries generated using terms are not as natural as those consisting of sentences.

In this paper, our proposed SumView system can automatically select the most representative sentences for each feature as the summary. Summaries generated by sentences are more intuitive and readable, thus it is easier for users to grasp their semantic meanings. Note that we do not predict the polarities of the opinion sentences for the features because polarities are too subjective for different users, and we believe that the text descriptions of features are more readable and helpful for users.

2.2. Multi-document summarization

This work is related to the problem of multi-document summarization. Here, we briefly discuss the related work on summarization.

- ***Centroid-based methods***: This type of methods ranks sentences by computing their salience using a set of features. For example, MEAD (Radev, Jing, Stys, & Tam, 2004) is a typical centroid-based algorithm which extracts sentences according to three parameters, i.e. centroid value, positional value, and first-sentence overlap. The centroid value of a sentence is computed as the average cosine similarity between the sentences and the rest of the sentences in the document collection. The positional value is computed as follows: the leading sentence is assigned score 1 and the score decreases by $1/n$ for each sentence, where n is the number of sentences in these documents. The overlap value is computed as the cosine similarity between a sentence and the first sentence in the same document. Then the three values are linearly combined with equal weights.
- ***Graph-based methods***: This type of methods constructs a sentence graph, in which each node is a sentence in the document collection, and if the similarity between a pair of sentence is above a threshold or the sentences belong to the same document, there is an edge between the pair of sentences. The sentences are selected to form the summaries by voting from their neighbors. Erkan and Radev Erkan and Radev (2004) propose an algorithm called LexPageRank to compute the sentence importance based on the concept of eigenvector centrality (prestige) which has been successfully used in Google PageRank. Other graph-based summarization have been proposed in Mihalcea and Tarau (2005) and Wan and Yang (2008).
- ***Latent semantic analysis (LSA)***: Gong and Liu (2001) propose a method using latent semantic analysis (LSA) to select highly ranked sentences for summarization. The method first creates a term-sentence matrix, where each column represents the weighted term-frequency vector of a sentence in the set of documents. Then singular value decomposition (SVD) is used on the matrix to derive the latent semantic structure. The sentences with the greatest combined weights across all the important topics are included in the summaries.
- ***Non-negative matrix factorization (NMF)***: This type of methods conducts NMF on the sentence-term matrix to extract sentences with the highest probability in each topic. NMF can also be viewed as a clustering method, which has many nice properties and advantages (Li & Ding, 2006). Intuitively, this method clusters these sentences and chooses the most representative ones from each cluster to form the summary.
- ***Other methods***: Other methods include CRF-based summarization (Shen, Sun, Li, Yang, & Chen, 2007), and hidden Markov model (HMM) based method (Conroy & O'Leary, 2001). Some

query-based summarization systems are also proposed (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999; Wan, Yang, & Xiao, 2007). For example, Language Computer Corporation (LCC) (See: http://www-nlpir.nist.gov/projects/duc/pubs/), a DUC participant, that proposes a system combining the question–answering and summarization system and using *k*-nearest neighbor clustering based on cosine similarity for the sentence selection. There also exist some techniques which utilize domain knowledge to help document summarization (Lee, Jian, & Huang, 2005; Pedrycz & Rai, 2008).

Most of the existing methods are designed for traditional document summarization tasks and do not make use of the product features. In this paper, our method is based on NMF framework and each cluster (or topic) is associated with one product feature.

3. System framework

Fig. 1 demonstrates the overall architecture of the SumView system. First of all, we develop a crawler to obtain product reviews from Amazon.com. Once a product ID (which is the unique product number provided by Amazon.com for each product) is given, all the user reviews and comments for this product are downloaded. The reviews from Amazon.com are free texts, then we decompose each review into sentences, and a POS tagger is used to identify each word. After removing stop words, the term-sentence matrix is constructed in the preprocessing step where each row represents a term and each column represents a sentence. Product features are automatically extracted using a method similar to the method proposed in Hu and Liu (2004), and the top five features are recommended to the users. The users can select any or any combination of them. In the meanwhile, users can also input their desired features by inputting in the text field. Fig. 2 shows an example. Based on the features that user selected and the constructed term-sentence matrix, the proposed feature-based weighted non-negative matrix factorization algorithm is performed to group the sentences into feature relevant clusters. Finally, the sentence with the highest probability in each cluster is selected as the summary for each feature. Fig. 3 shows tan example summary of reviews for a rice cooker.

4. Methodology

4.1. Opinion feature extraction and selection

For each category of products, we recommend five product features to users which are automatically extracted from the review texts using a method similar to the techniques proposed in Hu and Liu (2004). The task involves the following processes.

- Part of speech (POS) analysis is performed on the sentences. Only nouns and noun phrases are included in the candidate set.
- The term frequency-inverse sentence frequency (tf-isf) of the candidate terms are computed based on the term-sentence matrix, and top 20 candidates with highest tf-isf scores are kept in the candidate set. We consider them as frequent candidates.
- If the selected candidate nouns have a nearby adjective, we treat it as a frequent opinion feature.
- Finally, top five frequent opinion features are selected based on their frequency, and recommended to users.

To improve the usability of the system, SumView also provides a text field for users to input their desired product features. Users can select any feature or any combination of features from the recommended list. They can also input specific features they are interested in.

4.2. Feature-based initialization for non-negative matrix factorization (FNMF)

We propose a feature-based weighted non-negative matrix factorization method (FNMF) to generate review summaries. This method is based on non-negative matrix factorization (NMF) framework, and takes product features into consideration. As a result, each topic is related to a feature in the factorization results. FNMF automatically groups review sentences into the feature-relevant clusters, and selects the most representative sentences in each topic as the review summary.

In general, the FNMF method is composed of the following steps:

- Feature relevant initialization: Instead of random initializing U and V in the NMF algorithm where $A = UV^T$, the decomposed term-topic matrix U is initialized by selecting the sentence columns in A with maximum counts of each feature term, and the sentence-topic matrix V is consequently initialized by $(U^TU)^{-1}U^TA$.
- Non-negative matrix factorization (NMF): The NMF algorithm is performed on the weighted term-sentence matrix with the feature relevant initialization. The derivation of the NMF algorithm follows from Lee and Seung (SPS Year).
- Summary generation: After convergence of the NMF algorithm, the sentence with the highest probability for each topic is extracted from the sentence-topic matrix V to form the final summary.

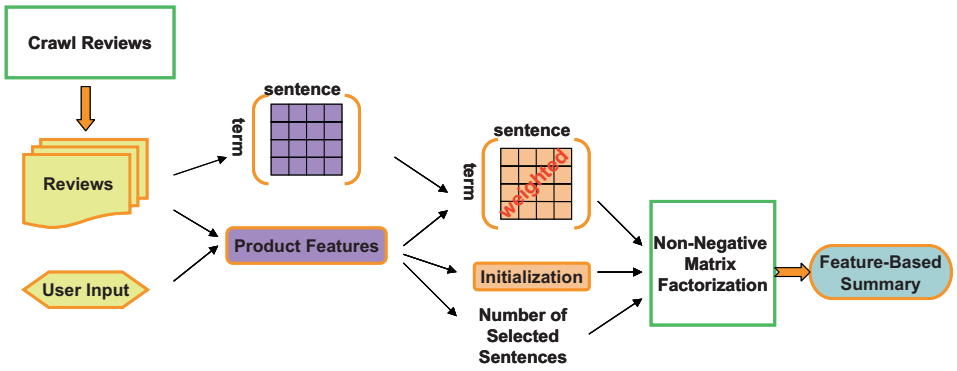

Fig. 1. The system overview.

Fig. 2. Product selection page.

Algorithm 1: Iterative Algorithm of FNMF

Input: A: term-sentence matrix
Output: U: term-topic matrix
V: sentence-topic matrix

1. initialize U by selecting the sentence columns in A with maximum counts of each feature term;
2. initialize $V = (U^TU)^{-1}U^TA$;
repeat
3. Update U using $U_{ij} = U_{ij}\frac{(AV)_{ij}}{(UV^TV)_{ij}}$;
4. Update V using $V_{ij} = V_{ij}\frac{(A^TU)_{ij}}{(VU^TU)_{ij}}$;
5. normalize U;
until convergence

The iterative algorithm is listed in Algorithm 1. Instead of selecting the most representative sentences in Step 3, an alternative approach for summary generation is to select several diverse sentences for each product feature so that different opinions can be shown to users. One feasible solution for selecting diverse sentences is that we can select a number of sentences with maximal relevance to the topic and minimum redundancy among these selected sentences. To implement this, we can sequentially select sentences one by one. Firstly the sentence with the highest probability for each topic is selected as described in Step 3, and then we select another sentence which has a high probability belonging to the topic but dissimilar to the selected sentences. Formally, we have

$$S_c = S_{c-1} \cup \left\{ max_{s_i \in X - S_{c-1}} (prob(s_i, topic) - \frac{\lambda}{c-1} \sum_{s_j \in S_{c-1}} sim(s_i, s_j)) \right\}, \quad (1)$$

where S_c is current selected sentence set, and S_{c-1} is previous set, and X is the complete sentence set. λ is a normalization parameter. In this paper, to examine the summarization ability and fairly compare with other summarization methods, we do not implement the diverse selection mechanism in the experiments.

5. Experiments

SumView is a summarization system which aims to summarize reviews based on product features, in which the feature identification approach applies the similar method in Hu and Liu (2004). Thus in this paper, our main contribution is how to utilize these features to obtain a meaningful feature-relevant summary for each product. Since existing review analysis systems usually focus on feature extraction and sentiment analysis, they do not have the similar functionality for generating natural summaries. Then, in the experiments, we compare our SumView with the following widely used document summarization systems to evaluate the summarization performance of SumView.

- *Centroid*: a centroid-based method ranking sentences based on their centroid values.
- *Graph*: a graph-based method performing PageRank on the sentence similarity graph.
- *NMF*: standard NMF for sentence clustering and selecting the sentences with highest probabilities in each topics.
- *Corpernic*: a commercial software for document summarization (Copernic, 2005).

5.1. A case study

First of all, we use a case study to demonstrate the review summaries generated by different summarization systems. In this case study, 173 reviews of a rice cooker from real customers have been crawled from Amazon.com (product ID: B000G30ESY). We choose this product because (1) a rice cooker is a good example of a home appliance that people use in their daily life; (2) the selected prod-

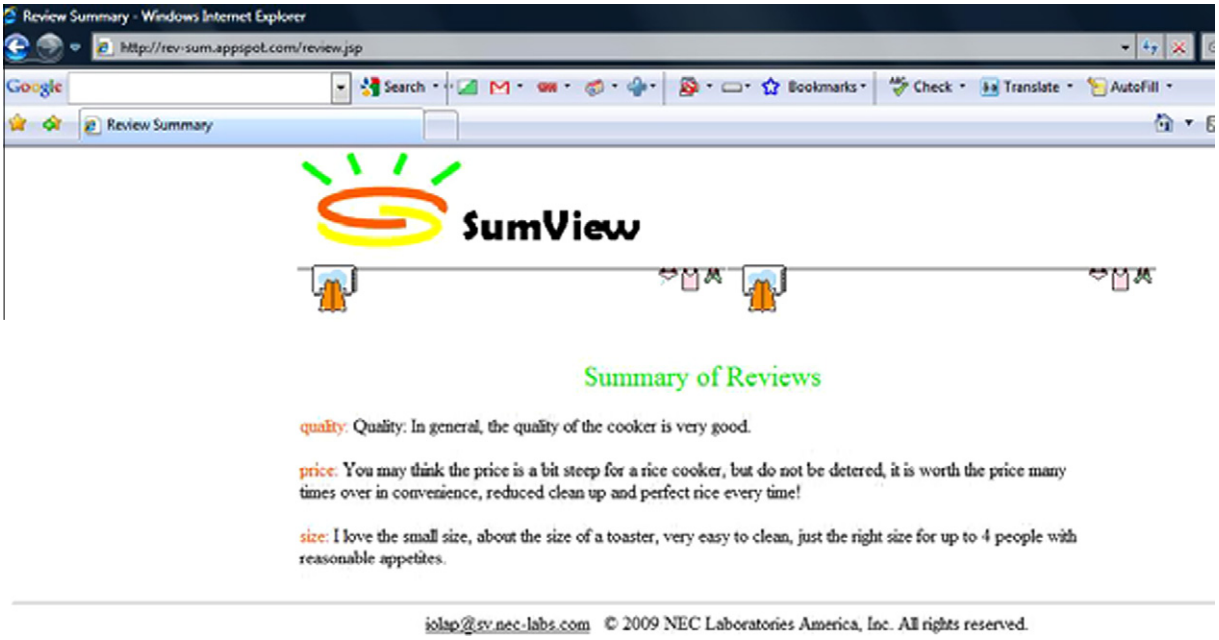

Fig. 3. An example summary.

uct is one of the best sellers. The top three product features extracted by the feature identification method described in Section 4.1 are "quality", "price", and "size". Table 1 shows the summaries generated from different summarization methods.

From the comparison of the generated summaries, we observe that:

- Centroid-based method has the worst performance. The results focus on the most frequent term "rice", and the selected sentences are redundant.
- Graph-based method capture the features of "size" and "price", and the 3rd sentence is also somehow related to "quality". However, all the three sentences do not reflect the majority of user opinions on these features. For example, the 2nd sentence is related to "price". In fact, there are 13 sentences in the review documents expressing users' opinions on the price of the cooker. 11 of them describe the cooker is worth buying on this price. Only two sentences express the price is too expensive, and one of which is the 2nd sentence selected by the graph-based method. In addition, the 3rd sentence is closely related to the "quality" issue, thus it is not representative.
- NMF clustering based method clearly outperforms the other methods. However, the first sentence is still not closely related to the cooking "quality".
- The summary generated by Copernic extracts the price and multifunction of the cooker. However, some information delivered by the summary is not focused on any product feature (e.g., the 3rd sentence of the summary).
- Comparing with the results using standard NMF which uses random initialization, we can see that the opinion on "quality" in the results of SumView is more focused and representative, which indicates the effectiveness of the feature-based initialization in our algorithm.

Note that another straightforward solution to use product features is to filter the sentences using these feature terms, i.e., only the sentences containing the feature terms are used in summarization. However, filtering sentences in this way will cause a big loss of information. From this case study, we clearly observe that The proposed FNMF algorithm in the SumView system can effectively utilize the feature information and outperform existing summarization techniques for results presentation.

Table 1
Different summarization methods based on the feature relevant weighted term-sentence matrix.

Centroid	• On my first try, the rice was even better than that! • It's just rice. • And, of course, that's the way to go for rice too.
Graph	• I take it right out of the freezer to use in other recipes, without defrosting it. • Anyway, this is still a nice cooker but not its price. • If the size is right, go for it.
NMF	• What's more, the rice comes out perfectly and evenly cooked. • You may think the price is a bit steep for a rice cooker, but do not be deterred, it is worth the price. • I love the small size, about the size of a toaster, very easy to clean, just the right size for up to four people with reasonable appetites.
Copernic	• Very good multicooker – has rice, porrige, slow cook, bake, steam and other modes. • I would highly recommend this to anyone else shopping for a rice cooker who would prefer not to spend a lot of money. • I bought this rice cooker to replace a 20 year old National rice cooker.
SumView	• In general, the quality of the cooker is very good. • You may think the price is a bit steep for a rice cooker, but do not be deterred, it is worth the price. • I love the small size, about the size of a toaster, very easy to clean, just the right size for up to four people with reasonable appetites.

Table 2
Description of the data sets.

	DUC2005	DUC2006
Number of topics	50	50
Number of documents relevant to each topic	25 ~ 50	25
Data source	TREC	AQUAINT corpus
Summary length	250 words	250 words

Table 3
Query-relevant multi-document summarization using DUC2005 data.

Systems	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-L	ROUGE-SU
Centroid	0.30573	0.04128	0.09421	0.26833	0.10121
Graph	0.31649	0.04960	0.11211	0.28973	0.10185
LSA	0.30461	0.04079	0.10883	0.26476	0.09352
NMF	0.31107	0.04932	0.10785	0.28716	0.10094
WFNMF	0.32812	0.05122	0.11637	0.30132	0.10538

Table 4
Query-relevant multi-document summarization using DUC2006 data.

Systems	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-L	ROUGE-SU
Centroid	0.32820	0.05285	0.11031	0.29862	0.10977
Graph	0.33046	0.05538	0.11743	0.31951	0.12113
LSA	0.33078	0.05022	0.11220	0.30507	0.10226
NMF	0.32374	0.05498	0.11341	0.31274	0.11348
WFNMF	0.34805	0.05731	0.12443	0.32318	0.12597

5.2. Summarization evaluation

In this set of experiments, we compare different summarization methods with our FNMF using Document Understanding Conference (DUC) benchmark data and Rouge evaluation toolkit (Lin & Hovy, 2003). Note that the DUC data consist of a number of collections of news articles and topic queries. Although the text content in news documents are different with that in product reviews, we still can capture term features from the topic queries (as keyword extraction) and then summarize the documents focusing on the extracted features. As there is no benchmark data and evaluation for review summarization tasks (especially for the task of selecting review sentences to form a short summary for a product), to obtain a subjective score-based performance evaluation, we use the topic-relevant multi-document summarization evaluation to compare the performance of our proposed feature-based initialization for NMF with other summarization methods.

5.2.1. DUC benchmark data sets

Table 2 describes the data used for summarization evaluation.

5.2.2. Rouge evaluation

We use ROUGE (Lin & Hovy, 2003) toolkit (version 1.5.5) to measure our proposed method, which is widely applied by DUC for performance evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries. Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-SU. ROUGE-N is an n-gram recall computed as follows

$$ROUGE-N = \frac{\sum_{S\in\{ref\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{ref\}}\sum_{gram_n\in S} Count(gram_n)}, \quad (2)$$

where n is the length of the n-gram, and ref is the reference summaries. $Count_{match}(gram_n)$ is the maximum number of n-grams co-occuring in a candidate summary and the reference summaries, and $Count(_{gramn})$ is the number of n-grams in the reference summaries. ROUGE-L uses the longest common subsequence (LCS) statistics, while ROUGE-W is based on weighted LCS and ROUGE-SU is based on skip-bigram plus unigram. Each of these evaluation methods in ROUGE can generate three scores (recall, precision and F-measure). As we have similar conclusions in terms of any of the three scores, for simplicity, in this paper, we only report the average F-measure scores generated by ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W and ROUGE-SU to compare our proposed method with other implemented systems.

Table 5
User satisfaction comparison.

Systems	SumView	Corpernic	Centoid	Graph	NMF	LSA	Term-RS	ListAll
User 1	4.2	3	2	2.4	2	2	3.2	2.2
User 2	3.6	2.2	1.6	1.6	1.4	1	2	1.6
User 3	3.4	1.6	1.6	1.8	1.8	1.4	2.6	1.2
User 4	4	2.6	2	2.2	2	2	2.4	2
User 5	4.6	2.2	2.2	2.6	2.6	2	3	2.2
User 6	3.2	1.4	1.6	1.4	1.4	1.4	1.4	2
User 7	3	1.4	1	1.6	1.6	1	1.4	1.4
User 8	3.8	2.2	1.4	1.8	1.6	1.4	2	2
User 9	3.4	1.6	1.6	2	2	1.6	2	2.2
User 10	4.4	2.8	2.6	3	2.8	2.6	3.2	2.4
User 11	3.4	1.4	1.8	1.4	1.6	1.6	1.6	1.8
User 12	3.2	1.6	1.2	1.8	1.2	1.4	2.2	1.4
User 13	3.8	2.2	1.8	2.4	2	1.8	3	1.8
User 14	4.4	2.8	2	2.8	2.8	2	3.6	1.6
User 15	3.8	2.2	1.6	2.4	2.2	1.6	2.2	1.4
User 16	3.6	2.6	1.8	2.6	2.2	2	2.4	2
User 17	3	1.4	1.6	1.6	1.8	1.6	1.8	2.2
User 18	4.4	2.4	2	2.8	2.2	2	2.4	2
User 19	3.8	2	2	2.2	2.2	2	2.8	1.8
User 20	4	2	1.6	2.4	2.6	1.6	3	2
User 21	3.2	2	1.4	2	2	1.6	2	3.8
User 22	3.2	1.4	1.6	2.2	2.2	2	3	2
User 23	4.4	3	2.2	3	2.8	2.2	3.8	2.4
User 24	3.4	1.8	2	2	2	1.4	2	3
User 25	3.8	2.8	2	3	3	2.6	4	2.8
Average	***3.72***	***2.10***	***1.87***	***2.2***	***2.08***	***1.75***	***2.52***	***1.97***

5.2.3. Experimental results

Tables 3 and 4 demonstrate the results by different summarization methods.

From the experimental results, we observe that the feature-based initialization in our WFNMF incorporates the features (keywords) analysis into document summarization effectively so that the performance of WFNMF outperforms traditional summarization methods and the standard NMF algorithm. One thing worth mentioning is that in this experiment we compare our method with the popular general summarization methods instead of some customized query-based summarization methods because those customized methods put more efforts on semantic analysis on query sentences which deviates the purpose of our experimental design. Thus, we only compare widely used general summarization methods with our method to demonstrate the summarization ability of our approach and the improvement of the feature-based initialization for the NMF algorithm.

5.3. User study

To better evaluate the results of SumView, we conduct a user survey. The subjects of the survey are 25 students at different levels and from various majors of a research university. First of all, we randomly select ten products from Amazon.com that have at least 30 reviews. Then each participant randomly selects five products from theses products and read their customer reviews. We generate review summaries using SumView and other baseline methods as described below. Then the participants are asked to assign a score of one (least satisfaction) to five (highest satisfaction), according to their satisfaction of the summaries generated by different methods. The baseline methods we use in this user study are:

- Traditional summarization methods including Centroid, Graph, LSA, and NMF.
- *Copernic*: a commercial document summarization software.
- *Term-RS*: a term-based review summarization system. Instead of selecting the most important opinion sentences, the adjectives for the opinion features are used as the summary of a product feature.
- *ListAll*: returning a list of opinion sentences without further summarization.

Table 5 presents the average satisfaction scores for each system from each participant.

We have found in this user study each product was selected by the users at least seven times and at most 18 times. And from the user assigned scores, we have the following observations. (1) Users do not satisfy with summaries generated by some traditional summarization methods such as Centroid and LSA, which indicates that an inappropriate summarization method can lead to worse user satisfaction than the results without summaries. (2) SumView outperforms all the traditional summarization methods because of the effectiveness of the feature extraction and feature-based initialization in our algorithm. (3) The fact that SumView outperforms existing term-based review summarization methods demonstrates the high usability of natural summaries generated by our system.

6. Conclusion

In this paper, we propose SumView, a Web-based review summarization system, to automatically extract the most representative expressions and customer opinions in the reviews on various product features. The system integrates review crawling, opinion product feature identification, and feature based sentence selection. The selected sentences represent the expressions and customer opinions in the product reviews on various product features. Comprehensive experiments and a case study demonstrate the effectiveness of the SumView system, and a user survey is also conducted to evaluate the users' satisfaction.

Acknowledgment

The project is partially supported by NSF Grants HRD-0833093, DMS-0915110, and CNS-1126619.

References

Archak, N., Ghose, A., & Ipeirotis, P. G. (2007). Show me the money!: Deriving the pricing power of product features by mining consumer reviews. In *Proceedings of SIGKDD*.

Carenini, G., Ng, R. T., & Pauls, A. (2006). Multi-document summarization of evaluative text. In *Proceedings of EACL*.

Cheung, K.-W., Kwok, J. T., Law, M. H., & Tsui, K.-C. (2003). Mining customer product ratings for personalized marketing. *Decision Support Systems, 35*(2), 231–243.

Conroy, J., & O'Leary, D. (2001). Text summarization via hidden markov models. In *Proceedings of SIGIR* (pp. 406–407).

Copernic. (2005). <http://www.copernic.com/en/products/summarizer/>.

Cui, H., Mittal, V., & Datar, M. (2006). Comparative experiments on sentiment classification for online product reviews. In *Proceedings of AAAI*.

Erkan, G., & Radev, D. (2004). Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Proceedings of the 6th international symposium on intelligent data analysis*.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of SIGIR* (pp. 121–128).

Gong, Y., & Liu, X. (2001). Generic text summarization usingrelevance measure and latent semantic analysis. In *Proceedings of SIGIR* (pp. 75–95).

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of SIGKDD* (pp. 168–177).

Jin, W., Ho, H. H. & Srihari, R. K. (2009). Opinionminer: A novel machine learning system for Web opinion mining and extraction. In *Proceedings of SIGKDD*.

Lee, D. & Seung, H. (2001). Algorithms for non-negative matrix factorization. In *NIPS*.

Lee, S., Jian, Z. W., & Huang, L. K. (2005). A fuzzy ontology and its application to news summarization. *IEEE Transactions on System, Man, and Cybernetics – PartB: Cybernetics, 35*(5), 859–880.

Li, T., & Ding, C. (2006). The relationships among various nonnegative matrix factorization methods for clustering. In *Proceedings of IEEE international conference on data mining* (pp. 362–371).

Li, T., Zhang, Y., & Sindhwani, V. (2009). A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. *Proceedings of the 47th Annual Meeting of the ACL, 1*, 244–252.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using *n*-gram co-occurrence statistics. In *Proceedings of NLT-NAACL* (pp. 71–78).

Lu, Y., Zhai, C., & Sundaresan, N. (2009). Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on world wide web: WWW '09* (pp. 131–140).

Meng, X., & Wang, H. (2009). Mining user reviews: From specification to summarization. In *Proceedings of ACL-IJCNLP*.

Miao, Q., Li, Q., & Dai, R. (2008). An integration strategy for mining product features and opinions. In *Proceedings of CIKM*.

Mihalcea, R., & Tarau, P. (2005). A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP 2005*.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM, 38*(11), 39–41.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.

Pedrycz, W., & Rai, P. (2008). A multifaceted perspective at data analysis: A study in collaborative intelligent agents. *IEEE Transactions on System, Man, and Cybernetics – PartB: Cybernetics, 38*(4), 1062–1072.

Popescu, A.-M., Nguyen, B., & Etzioni, O. (2005). Opine: Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP on interactive demonstrations*.

Radev, D., Jing, H., Stys, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, 919–938.

Shen, D., Sun, J.-T., Li, H., Yang, Q., & Chen, Z. (2007). Document summarization using conditional random fields. In *Proceedings of IJCAI* (pp. 2862–2867).

Somprasertsri, G., Lalitrojwong, P. (2008). A maximum entropy model for product feature extraction in online customer reviews. In *Proceedings of CIS*.

Tellis, G., & Johnson, J. (2007). The value of quality. *Marketing Science, 26*(6), 758–773.

Wan, X., & Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of the thirty-first annual international SIGIR conference*.

Wan, X., Yang, J., & Xiao, J. (2007). Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI* (pp. 2903–2908).

Yi, J. & Niblack, W. (2005). Sentiment mining in webfountain. In *Proceedings of ICDE*.